STATISTICAL ESTIMATION FROM AN OPTIMIZATION VIEWPOINT [†]

Roger J-B Wets

Department of Mathematics
University of California, Davis

Abstract. Statistics and Optimization have been closely linked from the very outset. The search for a 'best' estimator (least squares, maximum likelihood, etc.) certainly relies on optimization tools. On the other hand, Statistics has often provided the motivation for the development of algorithmic procedures for certain classes of optimization problems. However, it's only relatively recently, more specifically in connection with the development of an approximation and sampling theory for stochastic programming problems, that the full connection has come to light. This in turn suggests a more comprehensive approach to the formulation of statistical estimation questions. This expository paper reviews some of the features of this approach.

Key Words: constrained maximum likelihood estimation, consistency, epi-convergence

Date: June 30, 1998

[†] Research supported in part by a grant of the National Science Foundation

1. Introduction

From their inception Statistics and Optimization have been tightly linked. Finding the 'best' estimate of a statistical parameter is, of course, an optimization problem of some type. On the other hand, some of the motivation for the more recent development of solution procedures for *constrained* optimization problems came in part from statistical estimation questions. One can even trace back the development of the simplex method for linear programs to the work G.B. Dantzig did in his doctoral thesis on the Neyman-Pearson lemma [5], in particular, by underlining the importance of the 'column space' geometry when dealing with linear systems of equations and inequalities.

More recently, the design of approximation techniques and the need to justify the use of sampling to solve stochastic programming problems has revealed that besides the already widely accepted use of optimization techniques to find estimators, there is also a relationship between Statistics and Optimization at a more fundamental level. In turn, this has lead us to reappraise how one should formulate and deal with statistical estimation problems. In this discussion, we shall be mainly concerned with conceptual issues and will devote our attention to a few instructive examples. In the process, we review some of the theory that serves as the basis for this more comprehensive view of what is 'a statistical estimation problem'.

Most statistical estimation problems always can be formulated as finding the distribution function F of a random element (variable, vector, function, etc.). They come in two basic flavors. When prior information singles out a certain class of distribution functions characterized by a parameter $\theta \in I\!R^N$, then estimating F is reduced to finding a best estimate for this parameter. One refers to an estimation problem of this type as one of a *parametric estimation*. This means that the prior information available about the stochastic phenomenon described by the distribution function F is 'almost' complete, only the value to assign to some parameter(s) needs to be pinned down. When no prior information is available, the problem becomes one of *nonparametric estimation*, i.e., finding a function F whose only known property is that it is a distribution function.

These two problem types are in some sense at the opposite ends of what fits under the statistical estimation umbrella. Usually, some partial information is available about the unknown distribution, but not quite enough to be able to pinpoint the parametric class to which F belongs. For example, one might know, or suspect, that F is associated with a unimodal density function. One might know, or stipulate, bounds on certain quantiles. One might even know, or suspect that F belongs to a neighborhood of a certain distribution F^b (the Bayesian premise). In the same way that knowledge about the parametric class plays an important role in the (final) choice of the estimate, whatever information is available should be exploited in the choice of a 'best' estimate. In the formulation of the estimation problem, 'available information' gets translated in 'constraints' that restrict the choice of the estimate to a certain subclass of distribution functions.

We are going to refer to any (measurable) mapping from data to a space of parameters that identify distribution functions as an *estimator*. This space of parameters could be finite dimensional (parametric estimation) or infinite dimensional (nonparametric estimation). It could be the space of distribution functions itself.

To justify the choice of an estimator, one generally appeals to asymptotic analysis: one proves consistency [14] and, if possible, one derives a convergence rate that enables us to approximate the distribution of the error.

The basic example is the use of the sample mean as an estimator for the expected value of a real-valued random variable X (with finite mean μ and variance σ^2): Given ν samples $x^1, x^2, \dots, x^\nu$, the sample mean $(x^1, x^2, \dots, x^\nu) \mapsto \nu^{-1} \sum_{k=1}^{\nu} x^k$ is (asymptotically) consistent by the Law of Large Numbers. And, the Central Limit Theorem tells us that the distribution of the error $(\nu\sigma^2)^{1/2}\big(\nu^{-1} \sum_{k=1}^{\nu} x^k - \mu\big)$ tends to a standard Gaussian distribution.

Although asymptotic analysis has been the mainstay of Mathematical Statistics, when dealing with practical applications statisticians have often made use of estimators that might not (yet) have been subjected to full asymptotic analysis. One of the reasons for this technological gap between theory and practice is that to carry out the asymptotic analysis, the estimator should be 'nice': simple, smooth, ... And practice might suggest or dictate a choice which doesn't quite fulfill these requirements. In particular, this is usually the case when the estimator is the *argmax function*, i.e., a mapping which associates with a sample $x^1, \dots, x^\nu$ the solution of an optimization problem (based on these observations). This is exactly the approach that is going to be followed here. The estimation problem will always be formulated as an optimization problem:

find a distribution function $\widehat{F}$
that maximizes the likelihood of observing $x^1, \dots, x^\nu$
and so that $\widehat{F}$ is consistent with the available information.

Distributions functions $F : I\!R^d \to [0, 1]$ must already satisfy the constraints:

- F is nondecreasing, i.e., $x \le y \implies F(x) \le F(y)$;
- $\lim_{x_j \to \infty} F(x_1, \dots, x_j, \dots, x_d) = 1$, $\lim_{x_j \to -\infty} F(x) = 0$ for $j = 1, \dots, n$;
- F is upper semicontinuous;
- for all rectangles $R \subset I\!R^d$, $\Delta_R F \ge 0$ where $\Delta_R F$ is the measure assigned by F to R.

We won't deal here with the problem at this level of generality. In all of the examples and ensuing discussions, it will be assumed that it is known that the distribution of the random phenomenon is either discrete (the support is finite or countable) or continuous (there is a density function that determines the distribution function). But in all cases, the asymptotic analysis is based on the convergence of the argmax function of an optimization problem. In fact, it will be shown that the estimation problem converges to a certain limit estimation problem whose solution is the 'true' value of the parameter we are trying to estimate. The methodology is that of epi-convergence which is reviewed in the next section.

The choice of the maximum likelihood as the criterion isn't conditioned by the limitations of the methodology that's going to be used to obtain asymptotic justification; any other reasonable loss function will do equally well. However, one can argue, as already done by R.A. Fisher [7, 8], that the maximum likelihood is the 'natural' choice as criterion when selecting an estimator.

2. Epi-convergence: a primer

For the purposes of this paper, it will suffice to restrict our attention to functions defined on a separable Hilbert space $(H, |\cdot|)$. The usual framework in which this theory is presented [3, 12] is that of 'minimization'. The estimation problem being one of maximization, we should actually refer to it as *hypo-convergence*, see below, however it's expedient to keep the 'standard' name of 'epi-convergence' when refering to the general theory.

Let $f : H \to \overline{\mathbb{R}}$ be an extended real-valued function on H. Its *hypograph* is the set:

$$\operatorname{hypo} f := \{\, (x, \alpha) \in H \times \mathbb{R} \,|\, f(x) \geq \alpha \,\},$$

i.e., all the points in $H \times \mathbb{R}$ that lie on and below the graph of f; its *epigraph* is the set $\operatorname{epi} f := \{\, (x, \alpha) \in H \times \mathbb{R} \,|\, f(x) \leq \alpha \,\}$. Observe that f is *upper semicontinuous (usc)* if and only if $\operatorname{hypo} f$ is closed; recall that a function $f : H \to \overline{\mathbb{R}}$ is upper semicontinuous at x if $\limsup_{x' \to x} f(x') \leq f(x)$.

Definition. *A sequence $\{f^\nu : H \to \overline{\mathbb{R}},\ \nu \in \mathbb{N}\}$ hypo-converges to $f : H \to \overline{\mathbb{R}}$ at x, if*

$$\limsup_{\nu \to \infty} f^\nu(x^\nu) \leq f(x), \quad \forall\, x^\nu \to x;$$

and

$$\exists\, x^\nu \to x \ \ \textit{such that} \ \ \liminf_{\nu \to \infty} f^\nu(x^\nu) \geq f(x)$$

If this holds for all $x \in H$, the functions f^ν hypo-converge to f, f is called the hypo-limit of the f^ν, and one writes $f = \text{hypo-lim}_{\nu \to \infty} f^\nu$ or $f^\nu \xrightarrow{h} f$. The name *'hypo-convergence'* is motivated by the fact that this convergence notion is equivalent to the set-convergence of the hypographs.

Hypo-convergence yields the convergence of maximizers and optimal values, in a sense that will be made precise below, and it's all that's needed in many instances, in particular when H is finite dimensional. However, in infinite dimensions, it turns out that it is useful to introduce a somewhat stronger notion, namely *Mosco-hypo-convergence* which requires hypo-convergence with respect to both the weak and the strong topologies.

Definition. *A sequence $\{f^\nu : H \to \overline{I\!R},\ \nu \in I\!N\}$, with $(H, |\cdot|)$ a Hilbert space, Mosco-hypo-converges to $f : H \to \overline{I\!R}$ at x, if*

$$\textit{for all } x^\nu \underset{w}{\rightharpoonup} x \textit{ (weak convergence)}, \quad \limsup_{\nu\to\infty} f^\nu(x^\nu) \le f(x);$$

and

$$\exists\, x^\nu \to x \textit{ (strong convergence) such that } \liminf_{\nu\to\infty} f^\nu(x^\nu) \ge f(x).$$

If this is the case for all $x \in H$, the functions f^ν Mosco-hypo-converge to f, and one writes $f^\nu \xrightarrow{\text{M:h}} f$ or $f = M\text{:hypo-lim}_{\nu\to\infty} f^\nu$.

These two definitions should, more precisely, be qualified as 'sequential', but it won't be necessary to introduce this distinction here.

Theorem. *Suppose $\{\, f, f^\nu : H \to \overline{I\!R},\ \nu \in I\!N\}$ are such that $f^\nu \xrightarrow{h} f$, then*

$$\liminf_{\nu\to\infty} (\sup f^\nu) \;\ge\; \sup f.$$

Moreover, if there is a subsequence $\{\nu_k\}_{k\in\boldsymbol{N}}$, such that for all k, $x^k \in \operatorname{argmax} f^{\nu_k}$ and $x^k \to \bar{x}$, then $\bar{x} \in \operatorname{argmax} f$ and also $\sup f^{\nu_k} \to \sup f$.

If the functions f^ν Mosco-hypo-converge to f, and there is a subsequence $\{\nu_k\}_{k\in\boldsymbol{N}}$, such that for all k, $x^k \in \operatorname{argmax} f^{\nu_k}$ and $x^k \underset{w}{\rightharpoonup} \bar{x}$, then $\bar{x} \in \operatorname{argmax} f$ and $\sup f^{\nu_k} \to \sup f$.

Proof. These results are well known. We provide an elementary proof to illustrate the use made of the conditions in the definitions. The inequality $\liminf_\nu \sup f^\nu \ \ge\ \sup f$ certainly holds if $\sup f = -\infty$. If $\sup f$ is finite, then for all $\varepsilon > 0$ there exists x_ε such that $f(x_\varepsilon) > \sup f - \varepsilon$. Because the functions f^ν hypo-converge to f, there exists a sequence $x^\nu \to x_\varepsilon$ such that $\liminf_\nu f^\nu(x^\nu) \ge f(x_\varepsilon) > \sup f - \varepsilon$. This implies that $\liminf_\nu \sup f^\nu > \sup f - \varepsilon$, and since this holds for all $\varepsilon > 0$, it yields the desired equality. The case when $\sup f = \infty$ can be argued similarly, except that one now starts with the observation that for all κ there exists x_κ such that $f(x_\kappa) > \kappa$. The definition yields a sequence $x^\nu \to x_\kappa$ such that $\liminf_\nu f^\nu(x^\nu) \ge f(x_\kappa) > \kappa$, which again implies that $\liminf_\nu \sup f^\nu > \kappa$. Since this holds for all κ, it follows that $\liminf_\nu \sup f^\nu = \infty = \sup f$.

Now let $\{x^k,\ k \in I\!N\}$ be such that $x^k \in \operatorname{argmax} f^{\nu_k}$ for some subsequence $\{\nu_k\}_{k\in\boldsymbol{N}}$, and $x^k \to \bar{x}$. From the definition follows

$$\limsup\nolimits_k (\sup f^{\nu_k}) = \limsup\nolimits_k f^{\nu_k}(x^k) \le f(\bar{x}).$$

On the other hand,

$$\sup f \le \liminf\nolimits_k (\sup f^{\nu_k}) \le \limsup\nolimits_k (\sup f^{\nu_k}),$$

with the first inequality following from the argument above. Hence, $f(\bar{x}) = \sup f$, i.e., $\bar{x} \in \operatorname{argmax} f$. Moreover, this implies that the inequalities in the two preceding identities are actually equalities, and consequently $\sup f^{\nu_k} \to \sup f$.

In the case of Mosco-hypo-convergence, the argument is the same, except that the x^k converge weakly to $\bar{x}$ and one appeals to the corresponding condition in the definition of Mosco-hypo-convergence. $\square$

There are other consequences of hypo-convergence that are important in a statistical setting: for example, those characterizing hypo-convergence in terms of the convergence of the (super)level sets, or those dealing with convergence rates (metrics). But they won't be exploited here. For these results and more about the theory of epi-convergence, consult [3, 4, 12].

3. The discrete case

Let's identify the random phenomenon in which we are interested with a (generic) random variable X with values in $I\!R^d$. In this section, it will be assumed that this random variable takes on only a *finite* number of possible values, i.e., there is a finite collection of points $\{\, z^k \in I\!R^d, k \in K \subset I\!N \,\}$ with associated weights $\{\, p^0(z^k) \in (0,1), k \in K \,\}$ such that $\operatorname{prob}[\, X = z^k \,] = p^0(z^k)$; $p^0(z) = 0$ when $z \notin \{z^k,\, k \in K\}$. The distribution function of X is then

$$F^0(z) = \sum_{z^k \le z} p^0(z^k);$$

one refers to $p^0 : I\!R^d \to I\!R_+$ as the *probability mass function* of X. Its *support*, the smallest closed subset of $I\!R^d$ on which $p^0 > 0$, is the set $\operatorname{supp} p^0 := \{\, z^k \in I\!R^d, k \in K \,\}$ which is finite. This means that p^0 belongs to the following class of functions:

$$\operatorname{pmass}(I\!R^d; I\!R) = \{\, p \in \operatorname{fcns}(I\!R^d; I\!R) \mid \sum_{z \in \operatorname{supp} p} p(z) = 1,\ p \ge 0 \,\}.$$

The estimation problem is: given a sample $x^1, \dots, x^\nu$, obtained from independent observations, find $p^\nu \in \operatorname{pmass}(I\!R^d; I\!R)$ that approximates p^0 as well as possible. At this point, the only information available is the sample, and $\operatorname{supp} p^0$ is a finite set. In particular, this implies that $\{x^1, \dots, x^\nu\} \subset \operatorname{supp} p^0$.

One usually casts the estimation problem in the following mathematical framework: Let $\{X^k\}_{k=1}^\infty$ be *iid* (independent identically distributed) random variables, all with the same distribution as X, and for $j = 1, \dots, \nu$, let x^j be the observed value of X^j. The joint probability mass function of $X^1, \dots, X^\nu$ is given by

$$\operatorname{prob}\left[\, X^1 = z^1, X^2 = z^2, \dots, X^\nu = z^\nu \,\right] = p^0(z^1)p^0(z^2)\cdots p^0(z^\nu).$$

The probability of observing $x^1, \dots, x^\nu$ is then

$$p^0(x^1)p^0(x^2)\cdots p^0(x^\nu) = \prod_{l=1}^{\nu} p^0(x^l).$$

An estimate p^ν that maximizes the likelihood of observing $x^1, \dots, x^\nu$ is obtained as follows:

$$p^\nu \in \operatorname{argmax} \Big\{ \prod_{l=1}^{\nu} p(x^l) \,\Big|\, p \in \operatorname{pmass}(I\!R^d; I\!R) \Big\}.$$

After replacing the criterion function by its logarithm and setting $\ln 0 = -\infty$, the estimation problem can be reformulated as follows: find

$$p^\nu \in \operatorname{argmax} \Big\{ \sum_{l=1}^{\nu} \ln p(x^l) \,\Big|\, p \in \operatorname{pmass}(I\!R^d; I\!R) \Big\}.$$

Let's observe that only the values of p at $x^1, \dots, x^\nu$ play any role in the criterion function that is being maximized, consequently it never will pay off to choose a function p that assigns positive mass to any other points than $x^1, \dots, x^\nu$. Thus the problem of finding p^ν comes down to finding the values to assign to $p^\nu(x^1), \dots, p^\nu(x^\nu)$. Consequently, the estimation problem is equivalent to a finite dimensional optimization problem that we set up next.

Let $\{z^1, \dots, z^n\}$ be the distinct observations and r_i the number of times z^i has been observed. Let $(w_1^\nu, \dots, w_n^\nu)$ be the optimal solution of the following problem:

$$\max_{w \in I\!R^n} \quad L_c^\nu(w) := \begin{cases} \nu^{-1} \sum_{i=1}^n r_i \ln(w_i) & \text{if } w_i > 0,\ i = 1, \dots, n, \\ -\infty & \text{otherwise,} \end{cases}$$

$$\text{so that } \sum_{i=1}^n w_i = 1, \quad w_i \geq 0,\ i = 1, \dots, n.$$

It's the maximization of a continuous, strictly concave function on a nonempty compact subset of $I\!R^n$ (determined by linear constraints). This means that there is a unique optimal solution. The optimality conditions, cf. [12] for example, are

$(w_1^\nu, \dots, w_n^\nu)$ is optimal if and only if $\exists$ a multiplier $\theta^\nu \in I\!R$ such that
(a) $\sum_{i=1}^n w_i^\nu = 1, \quad w_i^\nu \geq 0,\ i = 1, \dots, n,$
(b) $(w_1^\nu, \dots, w_n^\nu) \in \operatorname{argmax}_{w \geq 0} [\nu^{-1} \sum_{i=1}^n r_i \ln(w_i) - \theta^\nu (\sum_{i=1}^n w_i - 1)].$

Setting the partial derivatives equal to 0 yields the (unique) solution:

$$w_1^\nu = \frac{r_1}{\nu},\ w_2^\nu = \frac{r_2}{\nu},\ \dots,\ w_n^\nu = \frac{r_n}{\nu}, \qquad \theta^\nu = 1.$$

The optimal estimate p^ν is then

$$p^\nu(z) = \begin{cases} \nu^{-1} r_i & \text{if } z = z^i, \\ 0 & \text{otherwise.} \end{cases}$$

This is the the *empirical probability mass function*; the probability mass assigned to each point $z \in I\!R^d$ is proportional to the number of times this point has been observed. The corresponding distribution function is the *empirical* distribution function F^ν with $F^\nu(z) = \sum_{z^k \le z} p^\nu(z^k)$. The support $\operatorname{supp} p^\nu$ of the empirical probability mass function consists of the points $z^1, \ldots, z^n$, observed so far. Thus,

> *processing (all) the statistical information available, i.e. the information provided by the sample, yields the empirical distribution as best estimate*
> No refinement of the estimate is possible *unless* additional information is available about F^0 or modeling assumptions are made about F^0.

This simply confirms our intuition! In the process it also provides support for the choice of the maximum likelihood as the criterion function. One implication is that we can formulate the estimation problem as one of maximizing an *expectation functional*:

$$\max_{p \in \mathrm{pmass}(\boldsymbol{R}^d;\boldsymbol{R})} L^\nu(p) = \int l(z,p)\, F^\nu(dz), \quad l(z,p) = \begin{cases} \ln p(z) & \text{when } p(z) > 0, \\ -\infty & \text{otherwise.} \end{cases}$$

It's convenient, although not literally correct, to refer to L^ν as the *maximum likelihood criterion.* The estimator is the argmax mapping:

$$F^\nu \mapsto p^\nu = \operatorname*{argmax}_{p \in \mathrm{pmass}(\boldsymbol{R}^d;\boldsymbol{R})} L^\nu(p);$$

the mapping which associates with the empirical measure F^ν the solution of the preceding optimization problem. In terms of our mathematical model, F^ν is a *random* distribution function since it depends on the values assumed by the random variables $X^1, \ldots, X^\nu$. In turn, this means that p^ν must also be viewed as a random variable. The asymptotic analysis of the argmax estimator has to do with the limiting properties of this random variable.

In this particular situation, because the empirical probability mass function is the optimal solution of the estimation problem, the Strong Law of Large Numbers —assuming $E\{|X_i|\} < \infty$— tells us that the p^ν converge almost surely to p^0, the 'true' probability mass function. Equivalently, the distribution functions F^ν converge almost surely to F^0. This shows that for the problem at hand, the argmax estimator is asymptotically consistent. Let's however proceed somewhat differently and derive consistency by appealing to the theorem in §2.

Let's begin by identifying the limit estimation problem:

$$\max_{p \in \mathrm{pmass}(\boldsymbol{R}^d;\boldsymbol{R})} L^0(p) = \int l(z,p)\, F^0(dz), \quad l(z,p) = \begin{cases} \ln p(z) & \text{when } p(z) > 0, \\ -\infty & \text{otherwise,} \end{cases}$$

where the expectation is now with respect to the true distribution function F^0. One might be tempted to justify the selection of this problem as the limit problem by appealing to the *Kullback-Leibler measure of discrepancy*, or even more simply by arguing that it's a natural choice since F^0 is the limit (in distribution) of the F^ν. However, in view of §2, to claim that this problem is the limit problem of the estimation problems, one needs to show that almost surely

$$\boxed{L^0 = \text{hypo-lim}_{\nu\to\infty} L^\nu \quad \text{on} \quad \text{pmass}(I\!\!R^d; I\!\!R)}$$

Note that p^0 is the optimal solution of this limit 'estimation' problem. Indeed, one can associate with this limit problem, the following optimization problem formulated in terms of the weights to assign to the points in the support of p^0:

$$\max_{w\in \boldsymbol{R}^{|K|}} L^0_a(w) := \begin{cases} \sum_{k\in K} p^0(z^k)\ln(w_k) & \text{if } \sum_{k\in K} w_k = 1,\ w_k > 0,\ k\in K, \\ -\infty & \text{otherwise,} \end{cases}$$

recall that $\text{supp}\, p^0 = \{\, z^k \in I\!\!R^d,\ k \in K \subset I\!\!N \,\}$, $|K|$ is the cardinality of K. The optimality conditions yield the following optimal solution: $w^0_k = p^0(z^k)$, $k\in K$, and this means that p^0 itself is the solution of the limit problem.

Instead of actually proving that almost surely $L^\nu \xrightarrow{h} L^0$ on $\text{pmass}(I\!\!R^d; I\!\!R)$, let's show that the following functions hypo-converge almost surely

$$L^\nu_a(w) := \begin{cases} \sum_{k\in K} p^\nu(z^k)\ln(w_k) & \text{if } \sum_{k\in K} w_k = 1,\ w_k > 0,\ k\in K, \\ -\infty & \text{otherwise,} \end{cases}$$

to L^0_a where all these functions are defined on $I\!\!R^{|K|}$. Operationally, in particular when actually computing the solution of the estimation problem, the functions L^ν_c defined earlier are used instead of L^ν_a; L^ν_c is just the restriction of L^ν_a to those points in $\text{supp}\, p^0$ that are included in the sample $x^1, \dots, x^\nu$.

By the Glivenko-Cantelli Theorem, p^ν converges uniformly to p^0 almost surely. And in turn, this implies that

$$L^\nu_a \text{ converge uniformly to } L^0_a \text{ almost surely;}$$

uniform convergence for extended real-valued functions being defined as follows: For a function $f: I\!\!R^n \to \overline{I\!\!R}$ and $\rho \in (0,\infty)$, let

$$f_{\wedge\rho}(x) = \begin{cases} -\rho & \text{if } f(x) \in (-\infty, -\rho), \\ f(x) & \text{if } f(x) \in [-\rho, \rho], \\ \rho & \text{if } f(x) \in (\rho, \infty). \end{cases}$$

A sequence of functions f^ν will be said to *converge uniformly* to f on a set $X \subset I\!\!R^n$ if, for every $\rho > 0$, their truncations $f^\nu_{\wedge\rho}$ converge uniformly to $f_{\wedge\rho}$ on X in the usual sense.

It remains only to observe that the almost sure uniform convergence of the upper semicontinuous functions L_a^ν to L_a^0 imply their almost sure hypo-convergence [12, Proposition 7.15].

Now, let $w^\nu = \operatorname{argmax} L_a^\nu$; recall that the strict concavity implies that the solution is unique. Since for all ν, $w^\nu \in \{\, w \in I\!R_+^{|K|} \mid \sum_{k\in K} w_k = 1\,\}$, a compact set, the sequence $\{w^\nu\}_{\nu\in I\!N}$ has at least one cluster point. And by the theorem in §2, every cluster point of this sequence must be the solution of the limit problem: $\max L_a^0$. The solution of this limit problem is also unique, viz., for $k \in K$, $w_k^0 = p^0(z^k)$. It follows that $w^\nu \to w^0$ almost surely. Translating this in terms of the corresponding probability mass functions yields $p^\nu \to p^0$ almost surely.

At this point one may wonder why this roundabout way of proving the consistency of the argmax-estimator! After all, a direct appeal to the Strong Law of Large Number already yielded the desired consistency. Moreover, the proof sketched out here does anyway appeal to the Law of Large Numbers via the Glivenko-Cantelli Theorem. There are two reasons for proceeding in this fashion. First, it's to set up the pattern that will be followed in other situations where it won't be possible to identify the estimate obtained any further than 'it's the solution of this optimization problem.' Second, it's to suggest that it's not just the estimates themselves but the entire estimation problems that converge, more precisely hypo-converge. Although it will not be possible to go into this in this short presentation, it allows for the study of the convergence of confidence regions and robustness properties as well.

But, let's now proceed to situations when more information is available about the distribution F, in which case the estimate generated by the argmax-estimator isn't in general the empirical distribution function.

p^0 **monotone:** Suppose the probability mass function p^0 is known to be nondecreasing, by which one means that $p^0(z^i) \le p^0(z^j)$ whenever $z^i \le z^j \in \operatorname{supp} p^0 \subset I\!R$. Our estimation problem should then be formulated as follows:

$$\max E^\nu\big\{\ln p(z)\big\} \text{ so that } \left\{ \begin{array}{l} \sum_z p(z) = 1,\ p(z) \ge 0, \\ \text{for all } z, z' \in \operatorname{supp} p,\ p(z) \le p(z') \text{ whenever } z \le z' \end{array} \right.$$

This is again a convex optimization problem. Because the solution will only assigns probability mass to the points $z^1, \ldots, z^n$ that have been observed, the problem can be recast as follows: Rearranging the observations so that $z^1 \le z^2 \le \cdots \le z^n$, with r_i be the number of times z^i has been observed and w_i the probability mass $p(z_i)$ to be assigned to z^i,

$$(w_1^\nu, \ldots, w_n^\nu) \in \operatorname{argmax}\{\sum_{i=1}^n \frac{r_i}{\nu} \ln w_i \,|\, 0 \le w_1 \le w_2 \cdots \le w_n, \sum_{i=1}^n w_i = 1\,\}.$$

The optimality conditions read: $(w_1^\nu, \ldots, w_n^\nu)$ is an optimal solution if and only if there are (multipliers) θ^ν, $\pi_1^\nu, \dots, \pi_{n-1}^\nu$ such that

$$\begin{aligned}
&\sum_{i=1}^n w_i^\nu = 1, \quad 0 \le w_1^\nu \le \cdots \le w_n^\nu,\\
&\pi_i^\nu \ge 0, \quad \pi_i^\nu(w_i^\nu - w_{i+1}^\nu) = 0, \quad i = 1, \dots, n-1,\\
&(w_1^\nu, \dots, w_n^\nu) \in \operatorname{argmax} \Big\{ \sum_{i=1}^n \frac{r_i}{\nu} \ln w_i - \theta^\nu (\sum_{i=1}^n w_i - 1) - \sum_{i=1}^{n-1} \pi_i^\nu (w_i - w_{i+1}) \Big\}.
\end{aligned}$$

From these conditions it follows that $(w_1^\nu, \dots, w_n^\nu)$ is the projection of the empirical weights $(r_1/\nu, \dots, r_n/\nu)$ on the convex polyhedral cone

$$M := \{\, (w_1, \dots, w_n) \in I\!R_+^n \,|\, 0 \le w_1 \le w_2 \le \cdots \le w_n \,\}.$$

The solution can be obtained recursively: set

$$\begin{aligned}
&q_i^1 = r_i/\nu, \quad i = 1, \dots, n,\\
&\begin{cases} \text{for all } i \text{, set } q_i^2 = q_i^1 \text{ if } q_2^1 \ge q_1^1,\\ \text{else } q_2^2 = q_1^2 = \frac{1}{2}(q_1^1 + q_2^1) \text{, set } q_i^2 = q_i^1 \text{, for } i > 2 \end{cases}\\
&\begin{cases} \text{for all } i \text{, set } q_i^3 = q_i^2 \text{ if } q_3^2 \ge q_2^2,\\ \text{else if } \frac{1}{2}(q_3^2 + q_2^2) \ge q_1^2 \text{, set } q_3^3 = q_2^3 = \frac{1}{2}(q_3^2 + q_2^2) \text{, for } i \ne 2, 3,\ q_i^3 = q_i^2,\\ \text{else } q_3^3 = q_2^3 = q_1^3 = (1/3)(q_3^2 + q_2^2 + q_1^2) \text{ and set } q_i^3 = q_i^2 \text{ for } i > 3, \end{cases}\\
&\dots \text{ and so on}
\end{aligned}$$

and finally, $(w_1^\nu, \dots, w_n^\nu) = (q_1^n, \dots, q_n^n)$.

To prove the consistency of this estimator a simple appeal to the Law of Large Number won't do. But the second argument based on the convergence of the argmax function goes through unchanged, except that the set of feasible probability weights is restricted to $\{\, w \in I\!R_+^n \,|\, \sum w_k = 1 \,\} \cap M$!

Clearly, a similar approach will work if the probability mass function is nonincreasing (on its support). The same recursive assignments will yield the optimal estimate $(w_1^\nu, \dots, w_n^\nu)$, except that one has to proceed in the reverse index order when projecting the empirical probability mass function onto the convex polyhedral cone

$$M' := \{\, (w_1, \dots, w_n) \in I\!R_+^n \,|\, w_1 \ge w_2 \ge \cdots \ge w_n \ge 0 \,\}.$$

Bounds on moments. If bounds are known for the kth moment of X, say $a_k \le E\{X^k\} \le b_k$, one includes a *linear* constraint in the formulation of the estimation problem. Again, with r_i the number of times z^i has been observed and w_i the probability mass to be assigned to z^i, our estimate $(w_1^\nu, \dots, w_n^\nu)$ is the solution of the following optimization problem:

$$\begin{aligned} \max \quad & \sum_{i=1}^n \frac{r_i}{\nu} \ln w_i \\ \text{so that } a_k \leq & \sum_{i=1}^n z_i^k w_i \leq b_k \\ & \sum_{i=1}^n w_i = 1, \quad w_i \geq 0,\ i = 1, \dots, n. \end{aligned}$$

Just as in the earlier cases, one can write down the optimality conditions, but they don't lead to a simple expression for the estimate. However, consistency can be argued exactly as earlier, except that this time the probability weights must also satisfy the constraint on the k-moment.

Bounds on the variance. Whereas bounds on the second moment lead to linear constraints, bounds on the variance might result in nonlinear constraints that are difficult to deal with. Indeed, if $v_l \leq \operatorname{var}(X) \leq v_u$, the constraints to be included in the estimation problem are:

$$v_l \leq \sum_{i=1}^n w_i (z_i)^2 - \Big(\sum_{i=1}^n w_i z_i \Big)^2 \leq v_u.$$

The lower bound constraint determines a convex region since the function

$$(w_1, \dots, w_n) \mapsto \sum_{i=1}^n w_i (z_i)^2 - \Big(\sum_{i=1}^n w_i z_i \Big)^2 \quad \text{is concave.}$$

But, the apparently more 'natural' constraint which comes from fixing an upper bound on the variance, leads to a nonconvex region. The implication being that the convex combination of two probability mass functions that both satisfy this constraint might very well not be acceptable. This should never be accepted without raising some questions about the formulation itself. Of course, the seasoned statistician might suggest that an upper bound on the variance should involve a dependence on the expectation and then the constraint would be of a different type. But, regardless of the convexity or of the lack of convexity resulting from this constraint, the consistency of the resulting argmax estimator is again guaranteed by the same argument as the one used in all the previous cases.

The examples considered are meant to illustrate both the general applicability of the consistency argument based on invoking hypo-convergence, but also the flexibility one has in the formulation of the estimation problem. It should be clear that one isn't limited to including just one additional constraint in the formulation of the estimation problem; for example, the case when the probability mass function is monotone involved a number of (linear) constraints.

It should also be pointed out that the constraints that have been introduced in the formulation of the estimation problem might play a significant role in determining the optimal estimate when the sample is relatively small. On the other hand, in a well-formulated estimation problem, these constraints should become inactive when the sample is quite large.

4. The parametric case

Again, let's identify the random phenomenon in which we are interested with a (generic) random variable X with values in $I\!R$. The case to be considered in this section is when the distribution of this random variable is known except for the value to assign to a finite number of parameters. We don't make a distinction between the cases when the distribution is discretely or continuously distributed.

Let $x^1, \dots, x^\nu$ be a sample coming from independent observations. The information available about X is this sample and the fact that the distribution function belong to a certain parametric class. The mathematical set up is the same as in §3: $\{X^k\}_{k=1}^\infty$ are *iid* (independent identically distributed) random variables all with the same distribution as X and for $j = 1, \dots, \nu$, x^j is then the observed value of X^j. As in §3, let $\{z^1, \dots, z^n\}$ be the collection of distinct observations from the sample $\{x^1, \dots, x^\nu\}$ and r_i the number of times z^i has been observed and the empirical measure assigns probability mass r_i/ν to z^i; expectation with respect to the empirical measure is denoted by $E^\nu\{\,\cdot\,\}$.

Our estimator is again the argmax mapping and the criterion is the maximization of the likelihood function; a brief justification for the use of the maximum likelihood as criterion was provided in §3 for the discrete case and §5 will deal with the continuous case.

X **is Poisson :** The probability mass function is

$$\text{prob}\,[\,X = k\,] = p(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, \dots .$$

Only the parameter λ needs to be determined. Adjusting the notation so that r_k denotes the number of times k has been observed, the estimation problem reads:

$$\max_{\lambda > 0} \quad \sum_{k=0}^{\infty} \frac{r_k}{\nu} \ln p(k) = \sum_{k=0}^{\infty} \frac{r_k}{\nu} \ln\left(e^{-\lambda}\frac{\lambda^k}{k!}\right);$$

no other constraints are needed here because $\lambda > 0$ implies $p^\nu(k) = e^{-\lambda^\nu}\frac{(\lambda^\nu)^k}{k!} \geq 0$ and $\sum_{k=0}^{\infty} p^\nu(k) = 1$. λ^ν is the estimate obtained by solving the optimization problem and p^ν is the resulting probability mass function.

Straightforward differentiation of the criterion function and setting the derivative equal to 0 yields

$$\lambda^\nu = \sum_{k=0}^{\infty} k\frac{r_k}{\nu},$$

i.e., the sample mean. To obtain the consistency of this estimator one could appeal directly to the Law of Large Numbers: the sample mean converges to the expectation of X. But one could also use the same argument as that used in all the examples in §3.

The same approach yields the well-known optimal estimates for the parameters of the binomial, Bernoulli, geometric, ..., probability mass functions, always under the assumption that no information is available beyond the parametric class and the sample. Also, the classical parametric estimators when X has a continuous distribution can be derived similarly. Let's go through a couple of examples.

X **exponentially distributed:** So, the density function of X is

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda e^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

Only the parameter λ needs to be determined. Because there are only a finite number of points in our sample, with probability one they are all distinct, so let's assume this to be the case: $(z^1, \dots, z^\nu) = (x^1, \dots, x^\nu)$, $r_i = 1$ for $i = 1, \dots, \nu$. The formulation of the estimation problem is then

$$\max_{\lambda \geq 0} \ \sum_{i=1}^{\nu} \frac{1}{\nu} \ln \left(\lambda e^{-\lambda z^i} \right).$$

Also here there is no need to impose any further constraints since every $\lambda > 0$ corresponds to an exponential density (that is nonnegative and sums up to 1). Differentiating yields $(\lambda^\nu)^{-1} = \sum_{i=1}^{n} (1/\nu) z^i$, i.e., the sample mean, and the corresponding (exponential) density is

$$f^\nu(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda^\nu e^{-\lambda^\nu x} & \text{if } x \geq 0. \end{cases}$$

X **normally distributed:** The density function of X is

$$f(x) = (\sigma^2 2\pi)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

Again, let's assume that the observations are distinct points. The estimation problem reads:

$$\max_{\mu, \sigma \geq 0} \ \sum_{i=1}^{\nu} \frac{1}{\nu} \ln \left(e^{-[\ln \sigma + \ln \sqrt{2\pi} + \frac{1}{2} \sigma^{-2} (z^i - \mu)^2]} \right),$$

or equivalently,

$$\max_{\mu, \sigma \geq 0} \ -\ln \sigma - \frac{\sigma^{-2}}{2\nu} \sum_{i=1}^{\nu} (z^i - \mu)^2.$$

Taking partial derivatives with respect to μ and σ obtains:

$$\mu^\nu = \frac{1}{\nu} \sum_{i=1}^{\nu} z^i, \quad \sigma^\nu = \left(\frac{1}{\nu} \sum_{i=1}^{\nu} (z^i - \mu^\nu)^2 \right)^{\frac{1}{2}}.$$

The argmax estimator for the pair (μ, σ) is, as expected, the vector made up of the sample mean and the sample's standard deviation.

The situation changes, for example, when there is an upper bound on the variance, say κ^2 $(\kappa > 0)$. The estimation problem reads:

$$\max_{\mu,\sigma} \; -\ln\sigma - \frac{\sigma^{-2}}{2\nu} \sum_{i=1}^{\nu} (z^i - \mu)^2 \text{ so that } 0 \le \sigma \le \kappa.$$

From the optimality conditions for this optimization problem, it's immediate that μ^ν is still the sample mean but

$$\sigma^\nu = \max\big[\,\kappa,\ \big(\frac{1}{\nu}\sum_{i=1}^{\nu}(z^i - \mu^\nu)^2\big)^{\frac{1}{2}}\,\big].$$

X **with known marginals:** Let's now consider the case in which $X = (X_1, X_2)$ is a 2-dimensional discrete random variable with known marginal for X_1, say

$$p_1(z^i) = \alpha_i, \quad i = 1, \ldots, n_1.$$

The estimation problem can be formulated as follows:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n_1}\sum_{i=2}^{n_2} \frac{r_{ij}}{\nu} \ln w_{ij} \\ \text{so that } & \sum_{j=1}^{n_2} w_{ij} = \alpha_i, \;\; i = 1, \ldots, n_1, \\ & w_{ij} \ge 0, \; i = 1, \ldots, n_1, \; j = 1, \ldots, n_2; \end{aligned}$$

r_{ij} is the number of times the pair (z^i, z^j) has been observed. It isn't necessary to include the constraint $\sum_{i=1}^{n_1}\sum_{i=2}^{n_2} w_{ij} = 1$, it would be redundant since $\sum_{i=1}^{n_1}\alpha_i = 1$. The optimality conditions for this problem are: for $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$,

(o) w^ν_{ij} determine the best estimate if there are (multipliers) π^ν_j, $j = 1, \ldots, n_1$ such that

(i) $\sum_{j=1}^{n_2} w^\nu_{ij} = \alpha_i$, $i = 1, \ldots, n_1$, and $w^\nu_{ij} \ge 0$.

(ii) $w^\nu \in \operatorname{argmax}\Big\{\, \sum_{i=1}^{n_1}\sum_{i=2}^{n_2} \frac{r_{ij}}{\nu} \ln w_{ij} + \sum_{i=1}^{n_1} \pi^\nu_i \big(\sum_{j=1}^{n_2} w_{ij} - \alpha_i\big) \Big\}$

And this yields

$$\pi^\nu_i = \frac{\sum_{j=1}^{n_2} r_{ij}}{\nu\alpha_i}, \qquad w^\nu_{ij} = \frac{r_{ij}\alpha_i}{\sum_{j=1}^{n_2} r_{ij}}.$$

The vector of probability weights w^ν determine the probability mass function p^ν, as in §3. As for all the preceding cases, consistency follows from the convergence of the argmax function under hypo-convergence; cf. the Theorem in §2 and the arguments used in the preceding examples, the only change is the definition of the region that identifies acceptable estimates.

As a last example in this section, let's consider the case when both X_1 and X_2 are binomial random variables but with unknown parameters θ_1 and θ_2 that need to be estimated. The argmax estimate is found by solving the following problem:

$$
\begin{aligned}
\max \quad & \sum_{i=1}^{n_1} \sum_{i=2}^{n_2} \frac{r_{ij}}{\nu} \ln w_{ij} \\
\text{so that } & \sum_{i=1}^{n_1} w_{ij} = \binom{k}{n_1} (\theta_1)^k (1-\theta_1)^{n_1-k}, \quad k = 0, \ldots, n_1, \\
& \sum_{j=1}^{n_2} w_{ij} = \binom{k}{n_2} (\theta_2)^k (1-\theta_2)^{n_2-k}, \quad k = 0, \ldots, n_2, \\
& \theta_1 > 0, \theta_2 > 0, \;\; w_{ij} \geq 0, \; i = 1, \ldots, n_1, \; j = 1, \ldots, n_2.
\end{aligned}
$$

It's not possible to obtain a closed form solution for this problem. For fixed (θ_1, θ_2), the problem has the structure of a so-called transportation problem and there is an efficient solution procedure for this class of problems. A decomposition procedure might actually work well here: the master problem concentrating on the search for the optimal (θ_1, θ_2) and the subproblem, which then has the structure of a transportation problem, being used to adjust the variables w_{ij} as needed.

Although the examples have remained relatively simple, they have once more demonstrated the flexibility of our approach in accommodating various levels of information with in all cases, hypo-convergence, the appropriate tool to deliver consistency.

5. The continuous case

The random phenomenon in which we are interested is still to be identified with a random variable X with values in $I\!R^d$. Here it's assumed that F^0, the distribution function of X, is continuous. Let $h^0 : I\!R^d \to I\!R_+$ be the associated density function. Again, $x^1, \ldots, x^\nu$ is a sample coming from independent observations. And at the outset, this sample and the fact that the distribution function is continuous is the total information available about X. The problem is to find a 'best' estimate for h^0.

The mathematical framework is the same as in §3: the sample is viewed as coming from observing iid random variables $\{X^k\}_{k=1}^\infty$ that have the same distribution as X. Because now any one point in the sample space has probability 0 of being observed, let's proceed with the assumption that all points observed are distinct. In particular this means that the empirical measure assigns probability mass ν^{-1} to each one of the points $x^1, \ldots, x^\nu$ in the sample.

Our estimator is again the argmax mapping and the criterion is the maximization of the likelihood function. A brief justification could go as follows: Since $X^1, \ldots, X^\nu$ are iid with density h^0,

$$
\text{prob}\left[\, X^1 \in I\!B_\infty(x^1, \delta), \ldots, X^\nu \in I\!B_\infty(x^\nu, \delta) \,\right] = \prod_{l=1}^{\nu} \int_{\boldsymbol{B}_\infty(x^l,\delta)} h^0(s)\, ds,
$$

where $\mathbb{B}_\infty(x,\delta) \subset \mathbb{R}^d$ is the ℓ_∞-ball centered at x of radius δ, i.e., a hypercube whose edges are of length 2δ. The likelihood of observing $x^1, \dots, x^\nu$, or more precisely points in their immediate neighborhoods, can be measured in the following terms:

$$\lim_{\delta \downarrow 0} \frac{1}{(2\delta)^d} \prod_{l=1}^{\nu} \int_{\boldsymbol{B}_\infty(x^l,\delta)} h^0(s)\, ds = \prod_{l=1}^{\nu} h^0(x^l).$$

The estimation problem can thus be formulated as:

$$\text{find} \quad h^\nu \in \operatorname{argmax} \Big\{ \prod_{l=1}^{\nu} h(x^l) \,\Big|\, \int_{\boldsymbol{R}^d} h(s)\, ds = 1, h \geq 0 \Big\},$$

or equivalently, after taking the logarithm of the criterion function: find

$$h^\nu \in \operatorname{argmax} \Big\{ \nu^{-1} \sum_{l=1}^{\nu} \ln h(x^l) \,\Big|\, \int_{\boldsymbol{R}^d} h(s)\, ds = 1, h \geq 0 \Big\},$$

or still,

$$h^\nu \in \operatorname*{argmax}_{h \in H} \Big\{ E^\nu\{\ln h(X)\} = \int \ln h(x)\, P^\nu(dx) \,\Big|\, \int_{\boldsymbol{R}^d} h(s)\, ds = 1, h \geq 0 \Big\},$$

where P^ν is the empirical measure and E^ν indicates that the expectation is with respect to the empirical measure P^ν. H is a function space, typically a subspace of $L^2(\mathbb{R}^d; \mathbb{R})$ with prescribed characteristics.

Even so, this problem isn't well defined. Indeed, it doesn't have a solution that would be considered to be a probability density function: h^ν turns out to be the summation of Dirac functions that assigns equal mass to each sample point, the counterpart of the empirical measure.

Especially in this situation, when we are dealing with the search of a function, an element of an infinite dimensional space, any 'prior' information, including modeling assumptions, that might be available will be extremely valuable when the samples at hand are too few to reach the asymptotic range, *as is almost always the case* when $d \geq 1$. Indeed, this 'prior' information constitutes a larger share of the total information available when only a small number of samples have been collected. In terms of the estimation problem, this means that the constraints that describe the prior information will, as they should, play a more significant role in determining the optimal estimate. For example, one might know, or suspect, that the density function h^0 is smooth, consequently, it would be natural to restrict the choice to a space of functions with prescribed smoothness properties. Or h^0 might be known to be unimodal and the search for an estimate might be restricted to functions of that type.

Smoothness. Let X be real-valued whose density function h^0 is 'smooth'. If the search for an estimate is restricted to differentiable functions, the following constraint could be included in the formulation of the estimation problem:

$$\int \frac{h'(x)^2}{h(x)}\,dx \le \beta.$$

The term on the left is called the Fisher information.

Bounds on moments. Suppose X is real-valued and there is some (prior) information about the expectation and the second moment of X, namely,

$$0 \le \mu_l \le E\{X\} \le \mu_u, \qquad \sigma_l^2 + \mu_u^2 \le E\{X^2\} \le \sigma_u^2 + \mu_l^2,$$

where we must have $\sigma_u^2 \ge \sigma_l^2 + (\mu_u^2 - \mu_l^2)$. This leads to introducing the constraints:

$$\mu_l \le \int x h(x)\,dx \le \mu_u, \qquad \sigma_l^2 + \mu_u^2 \le \int x^2 h(x)\,dx \le \sigma_u^2 + \mu_l^2$$

that imply that the sample variance must belong to $[\sigma_l^2, \sigma_u^2]$. By the way, observe that the set determined by this pair of constraints is convex.

Shape. Usually to include shape information an infinite number of constraints will be required. Assume $h^0 : I\!\!R^d \to I\!\!R_+$ is continuous and strongly unimodal, i.e., $h^0(x) = e^{-Q(x)}$ for some convex function Q. Assuming that Q is C^2, the constraints take the form

$$\langle z, \nabla^2 Q(x) z\rangle \ge 0, \qquad \forall\, z \in I\!\!R^d, \forall\, x \in I\!\!R^d$$

where $\nabla^2 Q(x)$ is the Hessian of Q at x. Here is another simple example of this ilk: consider the case when $h^0 : I\!\!R \to I\!\!R_+$ is known to be smooth and monotone decreasing. Then, the constraint $h'(x) \le 0$, for all $x \in I\!\!R$ should be included in the formulation of the estimation problem, cf. the numerical example at the end of this section.

Let A be the set of functions that satisfy the constraints generated by the 'prior' information. The consistency of the argmax estimator will be obtained, as in the previous sections, as a consequence of the consistency of the estimation problems:

$$h^\nu \in \operatorname*{argmax}_{h \in A \subset H} \Big\{ E^\nu\{\ln h(X)\} = \int \ln h(x)\, P^\nu(dx) \,\Big|\, \int_{\boldsymbol{R}^d} h(s)\,ds = 1, h \ge 0 \Big\},$$

with a limit problem whose solution is the true density function. Since the empirical measure P^ν can be viewed as approximating the probability distribution P of X, one might surmise (correctly) that this limit problem is:

$$\max_{h \in A \subset H} \Big\{ E\{\ln h(X)\} = \int \ln h(x)\, P(dx) \,\Big|\, \int_{\boldsymbol{R}^d} h(s)\,ds = 1, h \ge 0 \Big\},$$

where P is the probability measure associated with our random phenomena.

To conform to the framework of the hypo-convergence results in §2, let's identify the estimation problem with the function $E^\nu L : H \to \overline{\mathbb{R}}$ defined by

$$E^\nu L(h) = \int_{\boldsymbol{R}^d} L(x,h)\, P^\nu(dx),$$

and the limit optimization problem with the function $EL : H \to \overline{\mathbb{R}}$ defined by

$$EL(h) = \int_{\boldsymbol{R}^d} L(x,h)\, P^0(dx),$$

where

$$L(x,h) = \begin{cases} \ln h(x) & \text{if } h \in A \cap \{\, h \geq 0 \mid \int_{\boldsymbol{R}^d} h(x) = 1 \,\}; \\ -\infty & \text{otherwise.} \end{cases}$$

From the theorem in §2, the (Mosco-)hypo-convergence of the functions $E^\nu L$ to EL would follow the convergence of the estimates h^ν to h^0.

When dealing with consistency, however, one must take into account every possible sequence of samples, i.e., the $E^\nu L$ are actually random functions since they depend on the empirical measure P^ν that in turn depends on the observations of ν iid random variables $X^1, \ldots, X^\nu$. These random functions $E^\nu L$ are said to be *(strongly) hypo-consistent* with EL if they hypo-converge almost surely to EL. This is precisely what the Law of Large Numbers for random lsc functions guarantees.

As far as applications are concerned not much is lost if one picks H to be a separable reproducing kernel Hilbert space, typically H is one of the Sobolev spaces $H^p(\mathbb{R}^d)$ or $H^p_0(\mathbb{R}^d)$. So, let's proceed under this assumption.

Definition. *A function $f : \mathbb{R}^d \times H \to \overline{\mathbb{R}}$ is a random lower semicontinuous (random lsc) function if*

(i) for all $x \in \mathbb{R}^d$, the function $h \mapsto f(x,h)$ is lower semicontinuous,

(ii) $(x,h) \mapsto f(x,h)$ is $\mathcal{B}_d \otimes \mathcal{B}_H$-measurable, $\mathcal{B}_d, \mathcal{B}_H$ are the Borel fields on $\mathbb{R}^d$ and H.

Law of Large Numbers [2, 1, 11]**.** *Let H be a separable Hilbert space, $\{X^\nu,\ \nu \in \mathbb{N}\}$ a sequence of iid random variables with common distribution P on $\mathcal{B}_d$, and P^ν the (random) empirical measure induced by $X^1, \ldots, X^\nu$, and P^∞ the product measure on $\mathcal{B}^\infty_d$. Let $f : \mathbb{R}^d \times H \to \overline{\mathbb{R}}$ be a random lsc function and suppose that $\int \sup_h f(x,h)\, P(dx) > -\infty$. With $E^\nu f := \int_\Xi f(x,\cdot)\, P^\nu(dx)$ the (random) expectation of f with respect to P^ν, one has*

$$P^\infty\text{-almost surely:} \qquad \operatorname*{hypo-lim}_{\nu\to\infty} E^\nu f = Ef \qquad \text{where} \qquad Ef(h) = \int_{\boldsymbol{R}^d} f(x,h)\, P^0(dx).$$

Moreover, if P^0-almost surely, $f(x,\cdot) \leq \alpha_0 |\cdot - h'|^2 + \alpha_1(x)$ for some $h' \in H$, $\alpha_0 \in \mathbb{R}_+$, $\alpha_1 \in \mathcal{L}^1\big((\mathbb{R}^d, \mathcal{B}_d, P^0); \mathbb{R}\big)$, and $\int f(x, v(x))\, P^0(dx) > -\infty$ for some $v \in \mathcal{L}^2(\mathbb{R}^d; H)$. Then

$$P^\infty\text{-almost surely:} \qquad \operatorname*{Mosco-hypo-lim}_{\nu\to\infty} E^\nu f = Ef.$$

This Law of Large Numbers is the keystone to:

Hypo-consistency [10, 6]**.** *Let H be a separable reproducing kernel Hilbert space and $X^1, X^2, \ldots$, a sequence of iid $\mathbb{R}^d$-valued random variables. Suppose that*

$$S = \{\, h \in A \subset H \,|\, \int h(x)\, dx = 1,\ h \geq 0 \,\}$$

is closed. Then, the random lsc functions $E^\nu L$ hypo-converge P^∞-almost surely to EL with $E^\nu L$ and EL defined above, i.e., they are P^∞-a.s. hypo-consistent.

Under the additional conditions that S is convex, $x \mapsto \sup_{h\in S} h(x)$ is summable and $EL(h_*) > -\infty$ for some h_* such that $h_*^{1/2} \in H$, the random lsc functions $E^\nu L$ Mosco-hypo-converge P^∞-almost surely to EL.

There remains only to appeal to the theorem in §2 to obtain:

Consistency [6]**.** *Assuming that $h^0 \in A$ and under the same hypotheses as those yielding hypo-convergence, the estimates h^ν converge P^∞-almost surely to h^0. Under the additional assumptions that yield Mosco-hypo-convergence, for any weak-cluster point h^∞ of the sequence of estimates $\{\, h^\nu, \nu \in \mathbb{N} \,\}$ one has that $h^\infty = h^0$ P^∞-almost surely.*

The condition $h^0 \in A$ means that the constraints introduced in the formulation of the estimation problem do not exclude the 'true' density function. If h^0 doesn't belong to A, the estimates will converge to a density in A that is as close as possible, with respect to the Kullback-Leibler discrepancy measure, to h^0

A simple example will serve to illustrate the implementation of the overall strategy and, in particular, the role played by additional information. Let $x^1, x^2, \ldots, x^\nu$ be the sample generated from X, an exponentially distributed random variable with $E\{X\} = 1$. The information available: the sample, h^0 is smooth and decreasing.

To obtain the argmax-estimate h^ν one has to solve the infinite dimensional optimization problem:

$$\max_{h\in A\subset H} \sum_{\ell=1}^{\nu} \ln h(x^\ell) \ \text{ so that } \int_{\boldsymbol{R}} h(x)\, dx = 1,\ h \geq 0.$$

Such problems don't have a closed form solution. One has to resort to numerical techniques based on finite dimensional approximations. In [6], the approximation was built as follows: the Fourier series

$$\sqrt{1/\theta},\ \ \sqrt{2/\theta}\cos\Big(\frac{k\pi x}{\theta}\Big),\ k = 1, 2, \ldots,$$

is an orthonormal base for $H = \mathbb{L}^2([0,\theta]; \mathbb{R})$. Instead of $h \in H$, we consider only those functions in H obtained as a linear combination of the first q basis functions, viz.,

$$h_q(x,u) = u_0\sqrt{1/\theta} + \sum_{k=1}^{q} u_k \sqrt{2/\theta}\cos\Big(\frac{k\pi x}{\theta}\Big),$$

that depends only on the values assigned to the vector $u = (u_0, u_1, \dots, u_q)$. When the only information available about h^0 is the sample, the resulting nonlinear (semi-infinite) optimization problem becomes:

$$\begin{aligned}
\max \quad & \frac{1}{\nu}\sum_{\ell=1}^{\nu} \ln \Big[\frac{1}{\sqrt{\theta}} u_0 + \frac{\sqrt{2}}{\sqrt{\theta}} \sum_{k=1}^{q} \cos\Big(\frac{k\pi x^\ell}{\theta}\Big) u_k \Big] \\
\text{so that} \quad & u_0 + \sqrt{2}\sum_{k=1}^{q} \frac{\sin(k\pi)}{k\pi} u_k = 1/\sqrt{\theta}, \\
& u_0 + \sqrt{2}\sum_{k=1}^{q} \cos\Big(\frac{k\pi x}{\theta}\Big) \geq 0,\ \forall x \in [0,\theta], \\
& u_k \in I\!R,\ k = 0, \dots, q.
\end{aligned} \tag{UE}$$

Alternative formulations of this problems can be found in the monograph [13] of Thompson and Tapia on 'Nonparametric Function Estimation, Modeling and Simulation'.

If it's known that the density is a decreasing, more precisely nonincreasing, function on $[0, \theta]$, we need to include the constraints: $h_q(x) \geq h_q(x')$ whenever $x \leq x'$, and the estimation problem reads:

$$\begin{aligned}
\max \quad & \frac{1}{\nu}\sum_{\ell=1}^{\nu} \ln \Big[\frac{1}{\sqrt{\theta}} u_0 + \frac{\sqrt{2}}{\sqrt{\theta}} \sum_{k=1}^{q} \cos\Big(\frac{k\pi x^\ell}{\theta}\Big) u_k \Big] \\
\text{so that} \quad & u_0 + \sqrt{2}\sum_{k=1}^{q} \frac{\sin(k\pi)}{k\pi} u_k = 1/\sqrt{\theta}, \\
& u_0 + \sqrt{2}\sum_{k=1}^{q} \cos\Big(\frac{k\pi x}{\theta}\Big) u_k \geq 0, \quad \forall x \in I\!R, \\
& \sum_{k=1}^{q} \Big[\cos\Big(\frac{k\pi x}{\theta}\Big) - \cos\Big(\frac{k\pi x'}{\theta}\Big) \Big] u_k \geq 0, \qquad 0 \leq x \leq x' \leq \theta, \\
& u_k \in I\!R,\ k = 0, \dots, q.
\end{aligned} \tag{CE}$$

The difference between the estimates obtained via these two problems is illustrated by the following (typical) example: The sample size is 20 ($= \nu$), selected small on purpose. Any proposed nonparametric estimation method should work relatively well when the sample size is relatively large, but might fail to come up with believable results when the sample size is small. Kernel estimation techniques, for example, perform poorly when ν is small.

The solution of (UE) is graphed in Figure 1. The argmin-estimator was computed with $\theta = 4.1$ (substantially larger than any of the samples $x^1, \dots, x^{20}$) and $q = 3$, i.e., with four base functions; the mean square error was 0.02339. The use of a richer approximating

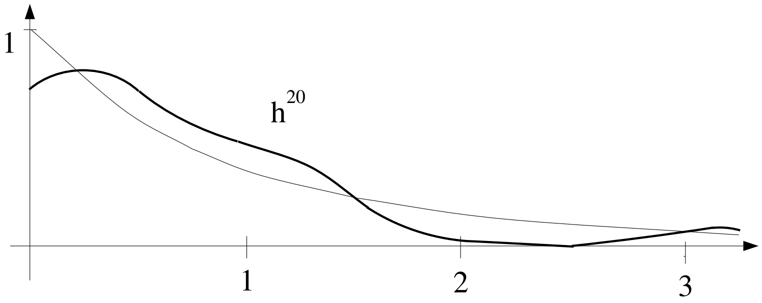

Fig. 1. argmin-estimator given 20 samples of an exponentially distributed random variable.

basis, i.e., with $q > 3$, yields estimates that oscillate in the tail and have slightly larger mean square errors.

The solution of (CE), that insists on densities that are nonincreasing, is graphed in Figure 2. Again with $\theta = 4.1$, but now $q = 5$, one obtains an estimate with mean square error of just 0.008743. Because more information is available in this case, one can reliably calculate the coefficients of a larger number of elements in the basis.

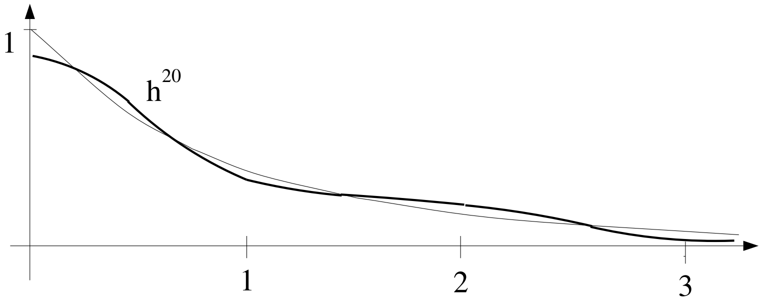

Fig. 2. argmin-estimator with monotonicity constraint.

It's informative to compare this approach via the argmax estimator to that followed by Groeneboom [9]. He suggested the following method for dealing with the estimation of nonincreasing density functions: Replace the empirical distribution function F^ν by G^ν that differs from it only on the interval I where $F^\nu \in (0,1)$, and on I, G^ν is the smallest concave function that majorizes F^ν. The resulting density function is a nonincreasing step function. One can prove consistency for this estimator which means that when the sample is increasing, the estimated density will eventually be arbitrarily close to h^0. However, whatever be the sample size one can never claim that the estimated density function is 'best' in any sense. In Groeneboom's approach, the use of the available information enters in the form of an afterthought, here it an intrinsic part of the calculation of the estimate.

Acknowledgement. I am grateful for the comments received from Lisa Korf (University of California, Davis) in the preparation of this article, and for Georg

Pflug's (University of Vienna) patience in enlightening me about statistical estimation and statisticians.

References

[1] Zvi Artstein & Roger J-B Wets, "Consistency of minimizers and the SLLN for stochastic programs," *J. of Convex Analysis* 1 (1995).

[2] Hedy Attouch & Roger J-B Wets, "Epigraphical processes: laws of large numbers for random lsc functions," manuscript University of California, Davis, 1991, (in Séminaire d'Analyse Convexe, Montpellier 1990, pp. 13.1–13.29).

[3] Jean-Pierre Aubin & Hélène Frankowska, *Set-Valued Analysis*, Birkhäuser Boston Inc., Cambridge, Mass., 1990.

[4] Gerald Beer, *Topologies on Closed and Closed Convex Sets*, Kluwer Academic Publishers, Dordrecht, 1993.

[5] George B. Dantzig & Abraham Wald, "On the fundamental lemma of Neyman and Pearson," *Annals of Mathematical Statistics* 22 (1951), 87–93.

[6] Michael X. Dong & Roger J-B Wets, "Estimating density functions: A constrained maximum likelihood approach," *Journal of Nonparametric Statistics* (1998), (to appear)..

[7] R.A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London, Series A* 222 (1922), 309–368.

[8] R.A. Fisher, "Theory of statistical estimation," *Proceedings of the Cambridge Philosophical Society* XXII (1925), 700–725.

[9] P. Groeneboom, "Estimating a monotone density," in *Proceedings of the Berkeley Conference in honor of J. Neyman and J. Kiefer* #2, Wadsworth Inc, 1985, 539–555.

[10] Christian Hess, "Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator," *The Annals of Statistics* 24 (1996), 1298–1315.

[11] Lisa Korf & Roger J-B Wets, "Ergodic limit laws for stochastic optimization problems," Manuscript, University of California, Davis, 1998.

[12] R.T. Rockafellar & Roger J-B Wets, *Variational Analysis*, Springer, Berlin, 1998.

[13] James R. Thompson & Richard A. Tapia, *Nonparametric Function Estimation, Modeling, and Simulation*, SIAM, Philadelphia, 1990.

[14] Abraham Wald, "Note on the consistency of the maximum likelihood estimate," *Annals of Mathematical Statistics* 20 (1949), 595–601.